(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 227 624 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
31.07.2002 Bulletin 2002/31

(51) Int Cl.[7]: **H04L 12/56**, H04L 29/06

(21) Application number: **01130529.9**

(22) Date of filing: **21.12.2001**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE TR
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **28.12.2000 US 750601**

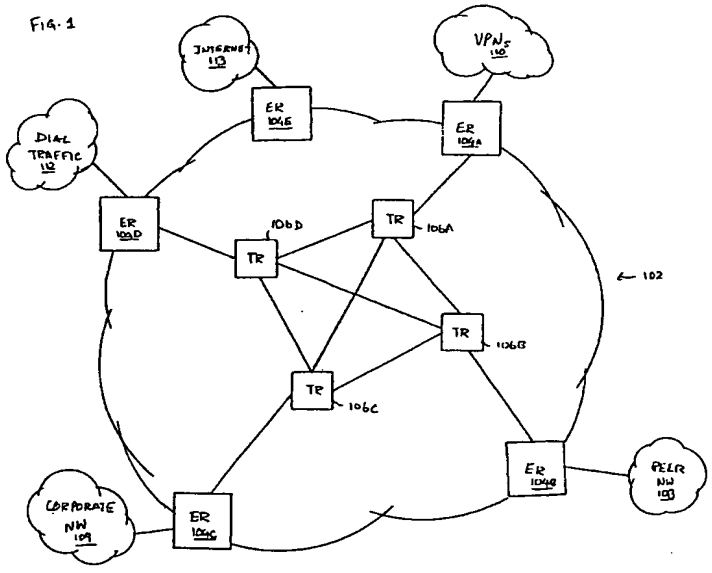
(71) Applicant: **Alcatel USA Sourcing, L.P.**
Plano, Texas 75075-5813 (US)

(72) Inventor: **Dharanikota, Sudheer**
Pleasanton, CA 94566 (US)

(74) Representative: **Schäfer, Wolfgang, Dipl.-Ing.**
Dreiss, Fuhlendorf, Steimle & Becker
Postfach 10 37 62
70032 Stuttgart (DE)

(54) Qos monitoring system and method for a high-speed diffserv-capable network element

(57) A QoS monitoring system and method for a Diff-Serv-capable network element operable in a trusted domain network such as an ISP network. The network element is organized as a plurality of terminating line cards interconnected via a switch fabric capable of supporting virtual ingress/egress pipes (VIEPs). Buffer queues on the ingress and egress sides of the network element, which are established for supporting traffic flows on individual VIEPS, are monitored for determining QoS parametric information such as throughput, loss, delay, jitter and available bandwidth. A policing structure is operably coupled with a buffer acceptance and flow control module for monitoring traffic behavior on the ingress side. Another buffer acceptance/flow control module and aggregate-level monitoring module are disposed on the egress side of the network element that cooperates with a scheduler which shapes outgoing traffic. The monitoring for the PIPE traffic reflects the conformance of the service provider to their customers, whereas the monitoring for the HOSE traffic reflects the level of over- or under-provisioning for a given COS. Feedback flow control is provided between the ingress and egress sides for throttling buffer acceptance.



EP 1 227 624 A2

SUMMARY OF THE INVENTION

[0009] Accordingly, the present invention provides a QoS monitoring system and method for a DiffServ-capable network element operable in a trusted domain network (such as an ISP/IAP network) that advantageously overcomes these and other shortcomings of the state-of-the solutions. Preferably, the trusted domain network is operable as an autonomous system wherein QoS parametric information may be monitored on multiple aggregate levels for SLA analysis, compliance and enforcement.

[0010] In one aspect, the present invention is directed to a network element (e.g., an edge router, core router, or transit router, collectively, a routing element) that is organized as a plurality of terminating line cards or TLKs interconnected via a switch fabric capable of supporting virtual ingress/egress pipes (VIEPs) between transmitter cards (ingress cards) and receiver cards (egress cards). Each TLK card is operable to support one or more incoming or outgoing communication links with respect to the network element, depending on its configuration. At least a portion of the TLK cards are operable as the network element's ingress side. Similarly, a portion of the TLK cards are operable as the egress side of the network element. Buffer queues on the ingress and egress sides of the network element, which are established for supporting traffic flows on individual VIEPs, are monitored for determining QoS parametric information such as throughput, loss, delay, jitter and available bandwidth. A policing structure is associated with the ingress cards for monitoring and measuring incoming traffic on the incoming communications links against an expected traffic profile or behavior pattern associated with the incoming traffic. A buffer acceptance and flow control module is associated with each of the ingress and egress cards that operates to manage the traffic flows associated with the VIEPs through the switch fabric. Preferably, the traffic flows are operable to be effectuated with resource reservations allocated in the switch fabric depending on type of service (e.g., real time vs. non-real time), Class of Service, SLA-based traffic engineering (TE) policies/priorities, et cetera. A traffic shaping and scheduling module is operable with an aggregate-level monitoring module disposed on the egress cards for scheduling and shaping outgoing traffic on the outgoing communications links to the network element's neighboring nodes in the network. Feedback flow control is provided between the ingress and egress sides for throttling buffer acceptance and packet discarding based on buffer congestion thresholds established on the egress side.

[0011] In another aspect, the present invention is directed to a method for processing QoS parametric information in a network element operable in an IP network, wherein the network element includes at least one terminating line card operable as an ingress card supporting an incoming communications link, at least one terminating line card operable as an egress card supporting an outgoing communications link and a switch fabric disposed between the ingress and egress cards for supporting a plurality of VIEPs therebetween. Upon receiving incoming information packets on the incoming link of the network element, a determination is made in an ingress portion a network processor system disposed on the ingress card whether the incoming information packets pertains to an IP-based service. Responsive to the determining step, the incoming information packets are forwarded to an egress portion of the network processor system via the switch fabric. The packets are monitored for conformance with respect to the reserved VIEP resources to the destination TLK (i.e., egress card). The processed information packets are transmitted to the egress card via a select VIEP for routing the processed information on a target outgoing link to a neighbor in the network. The egress portion preferably includes an embedded processor operable to perform a plurality of IP-based QoS (IPQoS) monitoring operations and for processing the incoming information into processed information.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] A more complete understanding of the present invention may be had by reference to the following Detailed Description when taken in conjunction with the accompanying drawings wherein:

[0013] FIG. 1 depicts an exemplary autonomous system operating as a trusted domain for coupling with a plurality of networks, wherein network elements incorporating the teachings of the present invention are advantageously employed;

[0014] FIG. 2 depicts a functional block diagram of an exemplary network element provided in accordance with the teachings of the present invention for operating in an trusted domain;

[0015] FIG. 3 depicts a functional block diagram of a network processor subsystem used in a terminating line card (TLK) of the exemplary network element of the present invention;

[0016] FIG. 4 depicts a functional block diagram of packet flow in the exemplary network element;

[0017] FIG. 5 is a message flow diagram for effectuating IP QoS monitoring in the exemplary network element in accordance with the teachings of the present invention;

[0018] FIG. 6 depicts a functional block diagram of a QoS monitoring system for use in the exemplary network element in accordance with the teachings of the present invention;

[0019] FIG. 7 depicts exemplary color monitors used as a component in a DiffServ traffic conditioner provided in the exemplary network element in accordance with the teachings of the present invention;

[0020] FIGS. 8A - 8C depict various packet discarding mechanisms that may be utilized as a component in flow

tion. A management server (MS) 218 is accordingly provided as part of the network element 200 for coordinating and hosting these administrative functions.

[0031] The functionality of TLK cards includes termination of external communication links, IP/MPLS forwarding, termination of link-related protocols (e.g., PPP, Label Distribution Protocol or LDP, Resource Reservation Protocol or RSVP, etc.) and switch interface functions such as Segmentation and Reassembly (SAR), resequencing, etc. The RTS boards complement forwarding capabilities of the TLK cards by providing the capability to process routing protocol messages. Routing protocols such as BGP, OSPF, etc., are typically processed on a route server (RS) 213 on the RTS boards. Consequently, forwarding tables are calculated and distributed through the MPSR switch fabric 204 by the route server to all forwarding engines on the TLK cards of the network element. In addition, the RTS boards are used for relaying external management messages from any external interface to MS 218. Further, an interface (e.g., Gigabit Ethernet (GENET) interface) (not explicitly shown in FIG. 2) to an external "charging server" may be included in the RTS boards for effectuating pricing policies of an SLA.

[0032] The functionality of the TLKs and RTSs is primarily carried out by one or more network processor (NP) modules (e.g., reference numeral 214) in conjunction with an on-board controller (OBC) processor 212. Each NP module is preferably comprised of an ingress portion 215A and an egress portion 215B. As will be seen in greater detail hereinbelow, an embedded processor (EP) 216 provided in the egress portion 215B is primarily responsible for the processing of incoming packets having IP service options (including IPQoS monitoring requirements). Moreover, EP 216 is also operable to process signaling protocol packets (origination or destination) for both L2 and L3 layers (e.g., PPP and GENET at L2 and RSVP and LDP at L3).

[0033] In addition to interfacing with the overlay communication network 220, OBC processor 212 is responsible for MPSR interface control. The switch generic interface functionality of the TLK/RTS card is comprised of a traffic manager (TM), which may be provided as a separate TM module 216 or as an embedded functionality of OBC processor 212, and a switch component. TM functionality is primarily responsible for effectuating the interface between the MPSR switch and the TLK/RTS card at all levels: physical level, logical-protocol level, and BW management level.

[0034] QoS-aware or QoS monitoring software applications running on EP 216 and OBC 212 are operable to inter-communicate via a TCP/IP protocol stack of resident operating systems (OS). For example, information regarding BW allocation for VIEPs in the switch is preferably communicated from an RSVP application (which is an EP process) to OBC in order to properly configure a TLK's TM. Further, the software environment of both processors preferably includes appropriate drivers (e.g., Peripheral Component Interconnect (PCI) drivers, etc.) for additional functionality.

[0035] Referring now to FIG. 3, depicted therein is a functional block diagram of the network processor subsystem 214 in additional detail. A networking function (NF) 232 is responsible for packet processing functionalities such as, e.g., forwarding, filtering, scheduling, etc. In addition to processing information packets with IP options, EP 216 is also operable to perform control functions such as IP control protocol message processing, exception processing, table management, and executing link-related signaling protocols. Several functional interfaces are associated with the NP subsystem 214 for facilitating its networking and QoS-aware functionalities. An external link interface 236 is provided for supporting incoming or outgoing links (e.g., links 206 and 208 depicted in FIG. 1) with respect to the network element. When configured as a receiver for packet information emanating from transmitting neighbors, the TLK having the NP module 214 is operable as an ingress card disposed on the ingress side of the network element. In similar fashion, a TLK having the NP module 214 may be configured as an egress card disposed on the egress side of the network element when packet information is transmitted via the external link interface 236 to the neighboring receiver elements. The external link interface 236 is therefore operable to receive/transmit packets towards the PHY layer devices that can be configured to support Layer 2 protocols, e.g., Fast Ethernet (FENET), GENET, etc., with appropriate media access control (MAC).

[0036] A switch interface 238 is provided for transmitting to or receiving from the switch fabric various intra-node traffic flows that are managed for QoS assurance. In some exemplary embodiments, a redundant switch interface may also be provided for increased reliability. A control memory array 234 is interfaced with the NP module 214 for storing control information, e.g., forwarding information base or bases (FIB), QoS monitoring counters, etc. Preferably, the control memory array 234 may be comprised of both SRAM and DRAM.

[0037] Continuing to refer to FIG. 3, a data buffer memory 240 is interfaced to the NP module 214 for storing information packets at the egress before they are transmitted on the external link or towards the switch fabric. Preferably, the data buffer memory 240 is implemented with double data rate (DDR) DRAM modules. A PCI interface 242 is provided for connecting an external host processor 244 such as, e.g., OBC processor 212 depicted in FIG. 2. Those skilled in the art should recognize that this interface is primarily used for system initialization and interaction of the NP module 214 with board and system management functions.

[0038] FIG. 4 depicts a functional block diagram of IP packet flow in the exemplary network element of the present invention. On the ingress side, the header of received frames is first parsed by a hardware (HW) classifier 302. This process classifies the packet depending

monitoring (described hereinbelow in additional detail) also takes place at this juncture. Thereafter, the outgoing traffic is shaped and scheduled for transmission (reference numeral 416) on an outgoing link.

[0045] FIG. 6 depicts a functional block diagram of a QoS monitoring system 500 for use in the exemplary network element in accordance with the teachings of the present invention. For purposes of ensuring DiffServ capability and corresponding SLA-based service constraints, various resource-based parametric monitors are advantageously employed as part of the QoS monitoring system of the present invention. For example, parametric information such as average occupancy of buffer queues, average over- and under-utilization of BW, etc. is deployed in order to manage appropriate aggregate-level QoS metrics. In a presently preferred exemplary embodiment of the present invention, these metrics include throughput, loss, delay, jitter, and available BW.

[0046] Throughput is defined as the average amount of data (in bytes and packets) transferred to a destination point per CoS. This measure is utilized to set up, manage, and identify different thresholds on the bulk of traffic flow per CoS. It should be appreciated that by employing per flow, per threshold levels, this measure becomes critically useful for effectuating a proactive action on the traffic behavior. Loss may be defined as the ratio of the amount of data dropped (in bytes and packets) to the amount of data transferred to a destination point per CoS. Accordingly, this metric measures the behavior of the buffer queues allocated to a particular traffic flow against their current reservation (i.e., queue utilization). Further, this metric also identifies to some extent the dynamic behavior of the queues and assists in performing reactive actions on the traffic behavior.

[0047] Delay is measured as the queuing delay in the system for different types of behavior. In addition to instantaneous values, the average behavior of this parameter is also important which depends on the CoS type. The average buffer queue depth is computed as the average of the instantaneous depths of the queue taken over a period of time. Jitter is defined as the variation or variance of the queuing delay. Average queue occupancies may be measured in relation to jitter monitoring in order to arrive at better measurements for the resource behavior. Available BW is the unreserved BW, which is monitored per link for traffic engineering purposes.

[0048] In order to monitor these QoS parametrics, the present invention provides structures and techniques for measuring the traffic characteristics on the ingress side as well as the egress side of the network element. A policing structure 504 is provided in the ingress TLK 202A which accepts a plurality of flows 502 (having different types) in order to measure the incoming traffic against the expected behavior. Traffic entering the DiffServ domain, wherein the network elements are provided in accordance with the teachings of the present invention, needs to be classified for appropriate treatment inside the domain. It must either be pre-marked by the customer or marked at the edge router level on the service provider's side of the network demarcation point.

[0049] Classification of customer traffic by the service provider's edge router can be based on multiple criteria, ranging from the interworking of various priority schemes to application level analysis of traffic within the IP packets as set forth hereinabove. Traffic policing may be implemented using a classifier (for classifying the incoming traffic), a token bucket or similar mechanism (for monitoring entry traffic levels at each class), and markers (for identifying or downgrading non-compliant traffic). FIG. 9 depicts a policing mechanism for throughput at an ingress TLK of the exemplary network element. As shown in FIG. 9, throughput monitoring is effectuated by tracking the in-profile and out-of-profile measurements over a time period. Throughput measurements are plotted against time as a profile 802 and a threshold 804 is defined for separating the in-profile portion 806B from the out-of-profile portion 806A. It should be appreciated that such aggregate-level throughput measurements suffice because of the incoming traffic flows at wire-speed.

[0050] The policing function is preferably effectuated by the NP module at both HW and code levels. A plurality of policing filters are provided as part of the policing structure 504, wherein one or more policing actions per packet are available. Further, policing may also be performed in accordance with a three-color marker (TCM) (described hereinbelow in reference to FIG. 7). Additionally, the loss parameter is measured as a projection of the traffic profile from the previous nodes that generate the incoming traffic towards the network element.

[0051] Continuing to refer to FIG. 6, a buffer acceptance and flow control module (hereinafter, a flow controller) is provided on both ingress and egress TLK cards. For example, flow controller 506 is provided as part of the functionality of the NP module (which may be referred to as the UP ramp NP module) of the ingress TLK 202A. Similarly, flow controller 510 is provided as part of the NP module (the DOWN or DN ramp NP module) of the egress TLK 202B. A QoS-aware traffic shaper/scheduler 508 is operable in association with the flow controller 510 on the egress TLK 202B for appropriately loading the outgoing links in accordance with QoS-based policies and constraints.

[0052] To monitor the characteristics of the traffic on the ingress and egress sides, various counters are implemented in association with the QoS-aware modules described hereinabove. Counters 506 are provided for the ingress TLK that measure (i) packets and bytes transferred per egress TLK per queue type and (ii) packets and bytes dropped per egress TLK per queue type. In similar fashion, counters 512 are provided for the egress TLK for measuring (i) packets and bytes transmitted in the egress TLK per neighbor per queue (accounts for the throughput); (ii) packets and bytes

ule are provided as part of the internal flow control that operates at the granularity of VIEPs. The CAC module distributes reserved BW pipes per VIEP based on periodic negotiation among all contending VIEPs. The IDRFC module is responsible for allocating non-reserved BW, which is distributed fairly among contending VIEPs in accordance with a Need For Bandwidth (NFB) resolution mechanism (which is preferably based on the UP ramp's per-VIEP buffer occupancy/arrival rate statistics).

[0064] As part of the policing functionality of the QoS monitoring system, the incoming traffic flows are classified, differentiated, and then broadly categorized into RT and RT queues. Depending on classes, traffic differentiators, etc., resource provisioning (e.g., buffers, BW and the like) for the RT and NRT traffic is done in accordance with QoS-based policies and constraints. For example, in setting up an RT flow which is preferably modeled in the "PIPE" model (where entry point and exit point of customer traffic is known), the following actions are taken: (i) GBW is reserved for a particular target port on the egress TLK card; (ii) CAC module associated with the egress TLK reserves this GBW for the VIEP associated with the RT flow; (iii) CAC module associated with the ingress TLK reserves appropriate GBW for the corresponding VIEP; and (iv) the policer parameters per target are updated in the ingress side. The NRT traffic is provisioned both in the PIPE model as well as the "HOSE" model (entry point is known but the exit point may be indeterminate). For setting up the NRT flows, the egress side NP module is first configured with GBW, AW, CBS, and PIR parameters. Optionally, per-class thresholds may then be configured therefor. Also, both ingress side and egress side CAC modules may be configured to reserve appropriate GBW for the VIEP associated with the NRT flow.

[0065] The QoS monitoring module of the present invention is operable to measure the behavior of the traffic due to the various reservations in the switch fabric per VIEP between the ingress and egress forwarding engines. Thus, the functionality of the buffer acceptance/flow controller module on the ingress side involves managing the queue behavior dynamics in the context of the egress scheduling, incoming traffic flows, and BW consumption in the switch fabric. The monitoring for the PIPE traffic reflects the conformance of the service provider to their customers, whereas the monitoring for the HOSE traffic reflects the level of over- or under-provisioning for a given COS.

[0066] Referring now to FIG. 10, depicted therein is a functional block diagram of a flow controller system 900 for use in the exemplary network element in accordance with the teachings of the present invention. Ingress TLK card 202A and egress TLK card 202B are exemplified once again as the ingress and egress sides of the network element. Each side is provided with CAC (which is configurable via appropriate signaling messages from the neighboring nodes) and IDRFC modules for BW reservation, VIEP arbitration, and internal flow control. In the exemplary embodiment depicted in FIG. 10, IDRFC 906A and CAC 908A are associated with the ingress side TLK and, in similar fashion, IDRFC 906B and CAC 908B are associated with the egress side TLK.

[0067] As described in detail hereinabove, policer 504 of the ingress side TLK is operable in conjunction with a packet discard structure 902A in order to condition the incoming traffic 502 for classification, differentiation, and categorization. A local congestion indicator 904A is provided to be operable in conjunction with the policer and packet discard mechanisms. Multiple flows targeted to the egress side TLKs are set up as a plurality of queues 910A, both RT (reference numeral 912A) and NRT (reference numeral 914A), wherein data buffers are allocated to each queue based on a flow control algorithm. It should be appreciated that the plurality of queues on the ingress TLK are indexed in terms of per egress card and per flow type (e.g., RT vs NRT).

[0068] Similarly, a plurality of queues 910B are set up on the egress side TLK for the outgoing traffic 509 emanating from the network element. Preferably, at least eight queues per neighbor are set up in a presently preferred exemplary embodiment of the present invention. These queues are indexed in accordance to per target port, per CoS, and per flow type. Thus, reference numerals 912B and 914 refer to RT and NRT queues for the egress TLK 202B. A local congestion indicator 904B is provided to be operable in conjunction with the egress side discard structure 902B, which in turn is associated with traffic shaper and scheduler 508.

[0069] Two factors are predominant in the measurement of the traffic behavior on the ingress side due to flow control, namely, (i) the flow control between the TLKs and (ii) the queuing for the RT and NRT queues between the TLKs. A data buffer to a particular egress TLK is accepted if the egress TLK is not congested. This information is received as a feedback flow control signal 913 from the egress congestion indicator 904B to a target threshold comparator 911 disposed in the ingress TLK 202A. It should be appreciated that where the number of buffers in these queues is sufficiently low, the delay and jitter parameters may be neglected in some implementations. However, under these low buffer numbers, the measurement of throughput and loss parameters becomes more significant. As the QoS module is operable to allocate resources through the switch for RT and NRT queues, throughput measurements can be advantageously used to determine whether the allocation for these two types of traffic is sufficient, both in qualitative and quantitative terms. Where there is an interaction between the RT and NRT sources, such interactions may be due to the condition that switch resource allocation cannot determined because the traffic cannot be characterized (e.g., incapable of identifying high priority routing traffic) or the traffic characteristics cannot be determined a priori (traffic modeled on the HOSE model and where there is no signaling involved).

3. The network element operable in an IP network as set forth in claim 2, wherein each of said flow controllers is associated with at least one local congestion indicator, and further wherein each of said flow controllers operate in conjunction with a packet discarding mechanism for throttling flow on a particular VIEP

4. The network element operable in an IP network as set forth in claim 3, wherein said packet discarding mechanism associated with said flow controller on said ingress card is operable to be controlled at least in part by a feedback signal received from said at least one local congestion indicator disposed on said egress card.

5. The network element operable in an IP network as set forth in claim 3, wherein said plurality of buffer queues are categorized as one of real time (RT) and non-real time (NRT) queue types.

6. The network element operable in an IP network as set forth in claim 5, wherein said plurality of counters associated with said ingress card include a counter for monitoring the number of packets transferred per egress card per queue type.

7. The network element operable in an IP network as set forth in claim 5, wherein said plurality of counters associated with said ingress card include a counter for monitoring the number of bytes transferred per egress card per queue type.

8. The network element operable in an IP network as set forth in claim 5, wherein said plurality of counters associated with said ingress card include a counter for monitoring the number of packets dropped per egress card per queue type.

9. The network element operable in an IP network as set forth in claim 5, wherein said plurality of counters associated with said ingress card include a counter for monitoring the number of bytes dropped per egress card per queue type.

10. The network element operable in an IP network as set forth in claim 5, wherein said QoS parametric information comprises traffic flow throughput information per Class of Service (CoS) per destination.

11. The network element operable in an IP network as set forth in claim 5, wherein said QoS parametric information comprises traffic loss ratio per Class of Service (CoS) per destination.

12. The network element operable in an IP network as set forth in claim 5, wherein said QoS parametric information comprises queuing delay information.

13. The network element operable in an IP network as set forth in claim 12, wherein said queuing delay information comprises average depth of said buffer queues on said ingress and egress cards.

14. The network element operable in an IP network as set forth in claim 5, wherein said QoS parametric information comprises traffic flow jitter information.

15. The network element operable in an IP network as set forth in claim 14, wherein said wherein said QoS parametric information comprises available bandwidth per link.

16. The network element operable in an IP network as set forth in claim 5, wherein said plurality of counters associated with said egress card include a counter for monitoring the number of packets transmitted per egress card per queue type for each neighboring element associated with said network element in said IP network.

17. The network element operable in an IP network as set forth in claim 5, wherein said plurality of counters associated with said egress card include a counter for monitoring the number of bytes transmitted per egress card per queue type for each neighboring element associated with said network element in said IP network.

18. The network element operable in an IP network as set forth in claim 5, wherein said plurality of counters associated with said egress card include a counter for monitoring the number of packets dropped per egress card per queue type for each neighboring element associated with said network element in said IP network.

19. The network element operable in an IP network as set forth in claim 5, wherein said plurality of counters associated with said egress card include a counter for monitoring the number of bytes dropped per egress card per queue type for each neighboring element associated with said network element in said IP network.

20. The network element operable in an IP network as set forth in claim 5, wherein said plurality of counters associated with said egress card include a counter for monitoring average queue depth of said buffer queues per egress card.

21. The network element operable in an IP network as set forth in claim 5, wherein said plurality of counters associated with said egress card include a counter for monitoring the number of times a particular buffer queue crosses a packet discard threshold associated therewith.

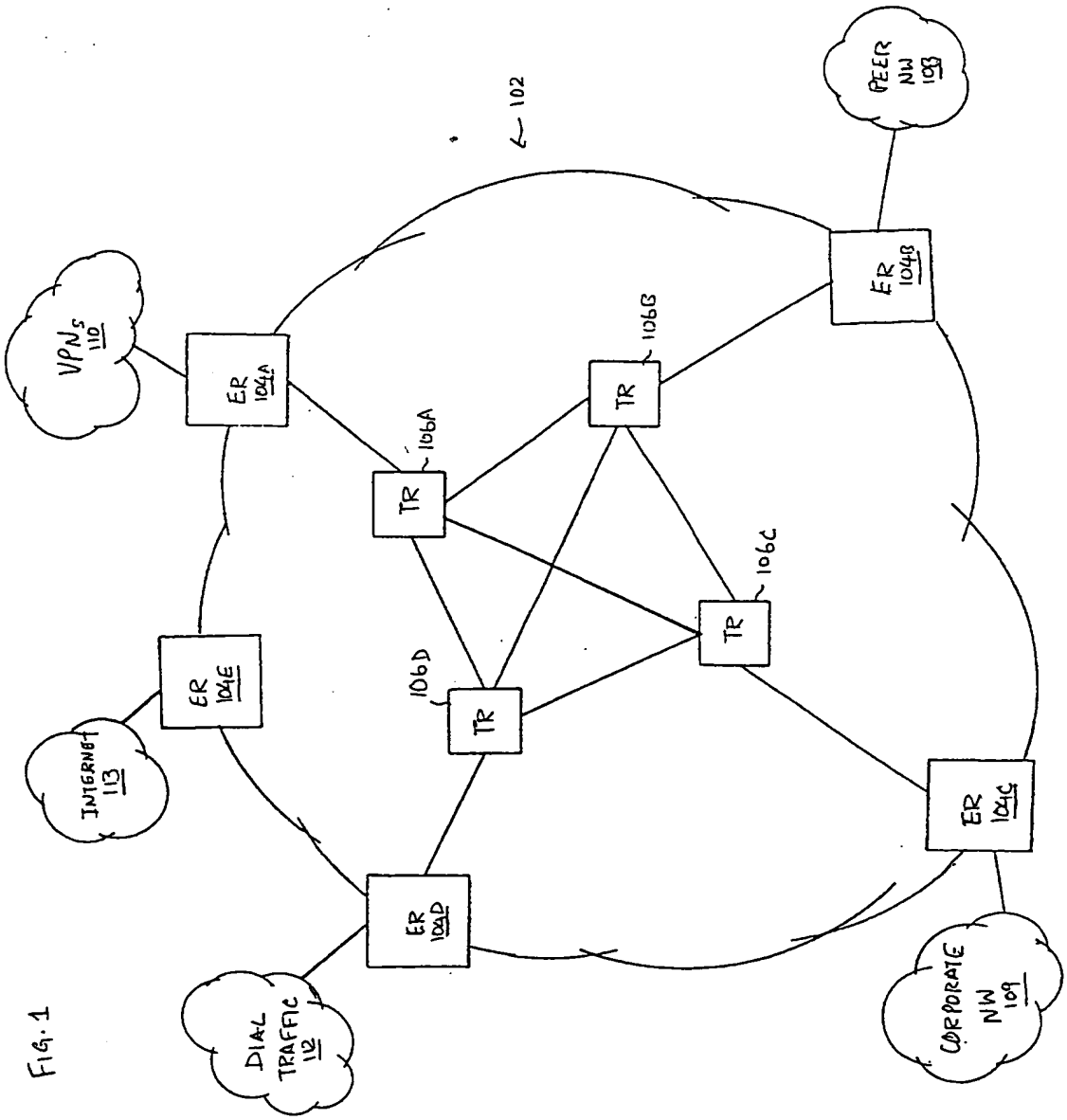


FIG. 1

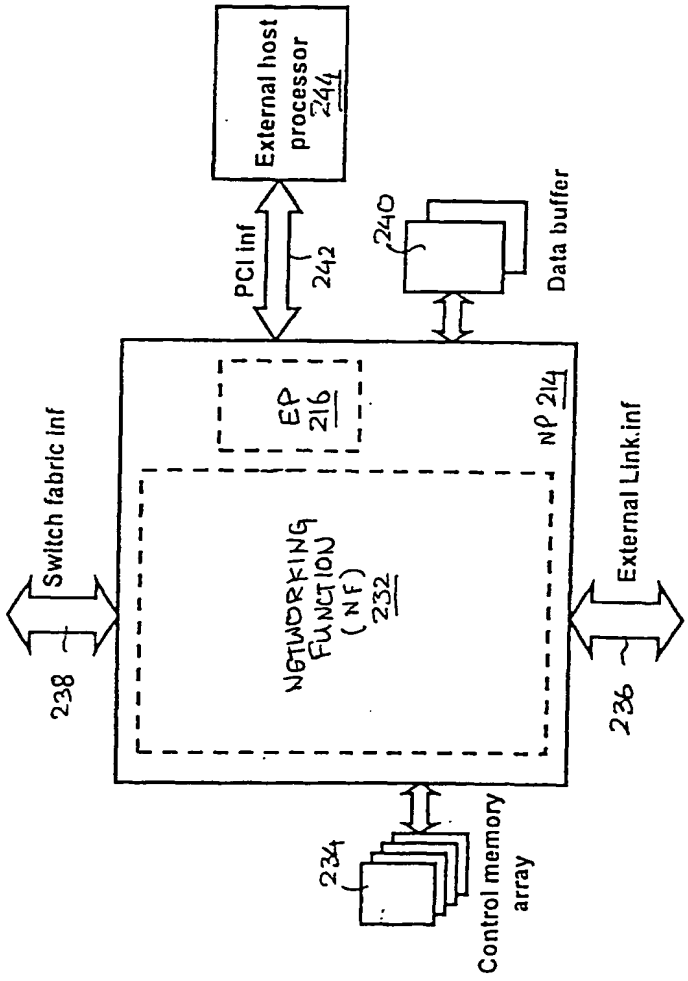
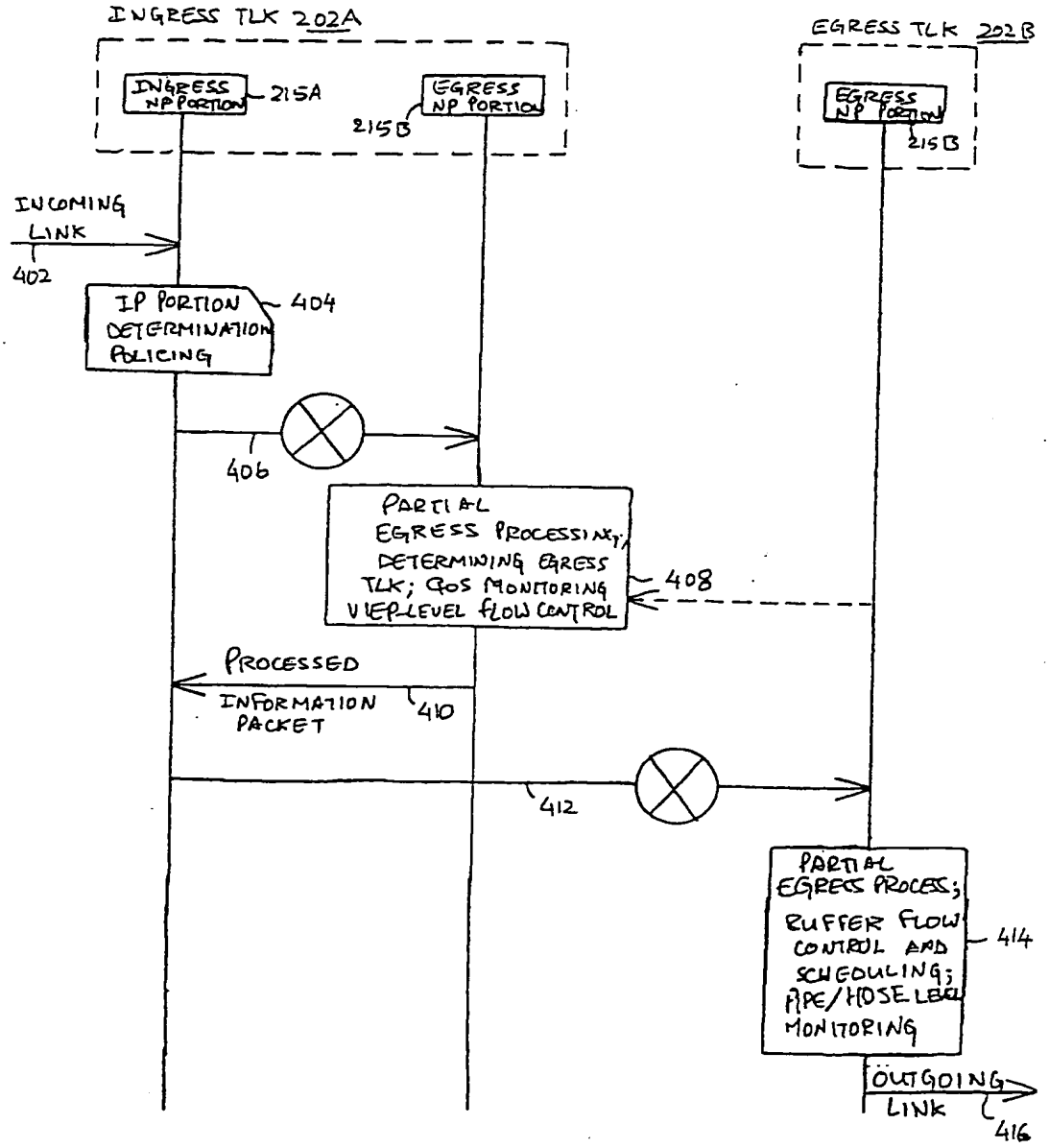


FIG. 3

FIG. 5



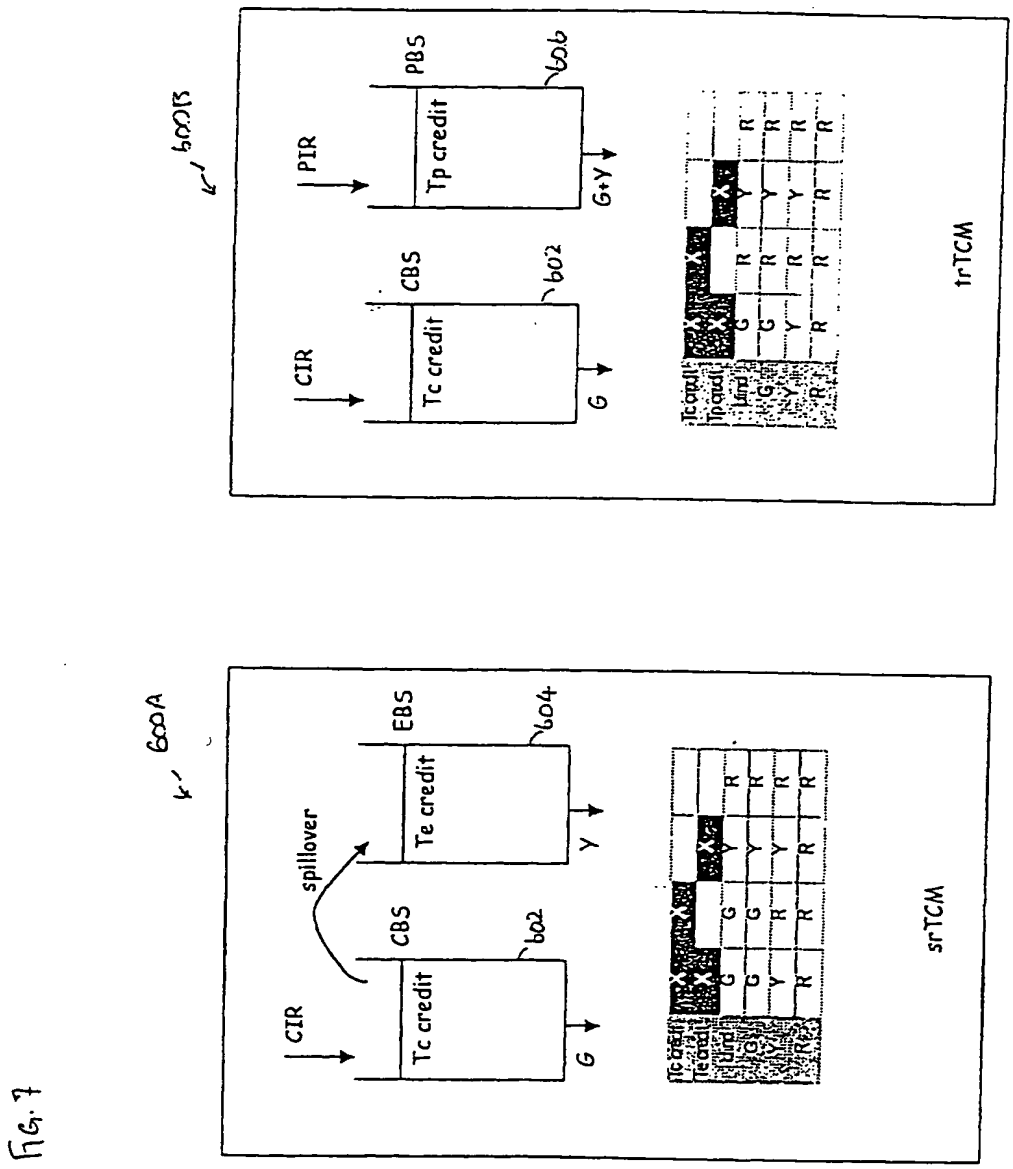


Fig. 7

FIG 8B

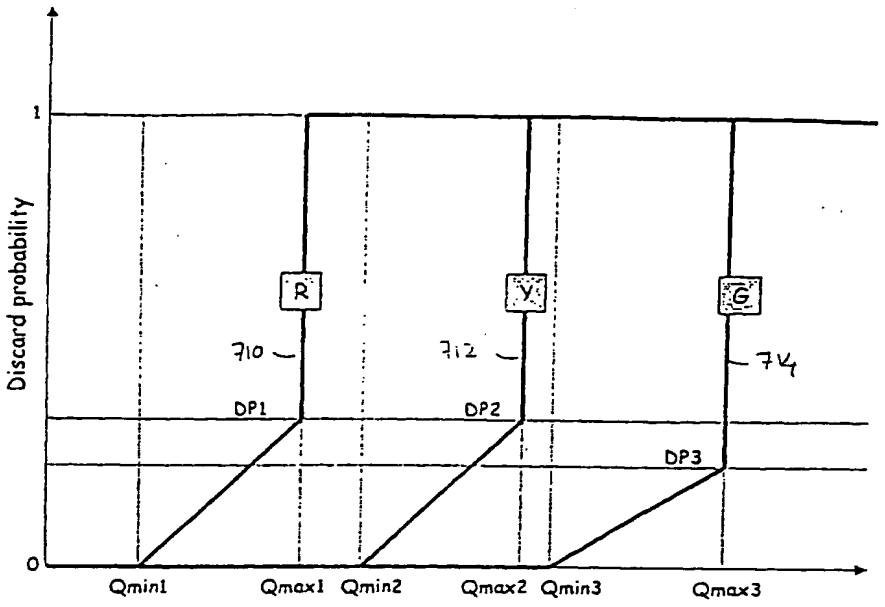
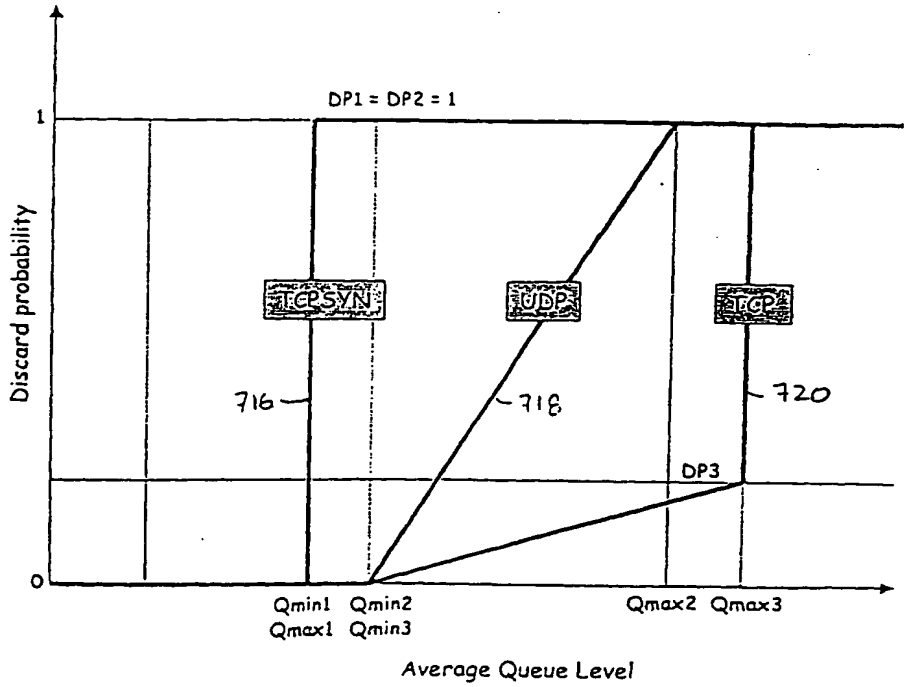


FIG 8C



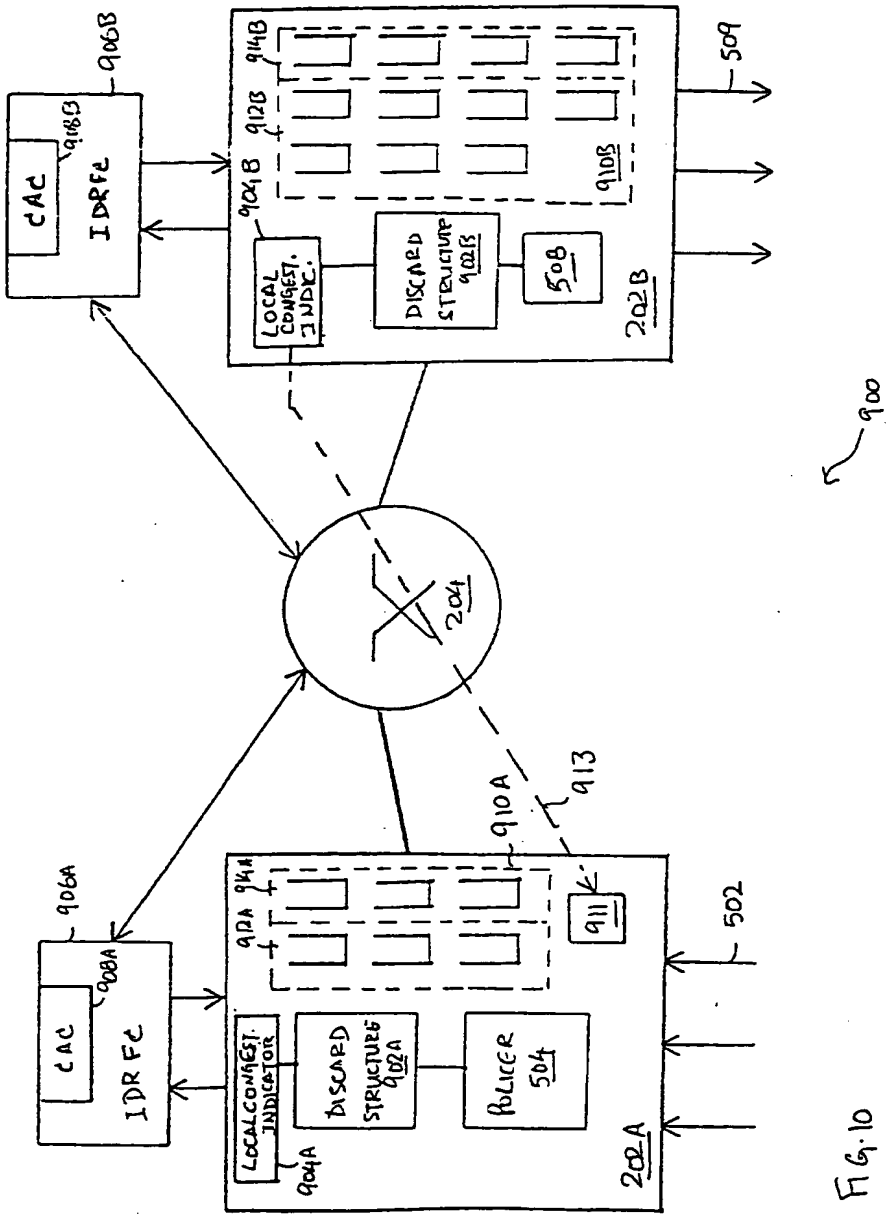


FIG. 10